Importing the Libraries

In [1]:

```
import json
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np


with open('brands.json') as f:
    brands_data = [json.loads(line) for line in f]


brands_flat = pd.json_normalize(brands_data)

brands_df = pd.DataFrame(brands_flat)
```

In [2]:

```

# Checking for duplicate records
duplicates = brands_df.duplicated().sum()
print(f"Number of duplicate records: {duplicates}")
```

```
Number of duplicate records: 0
```

In [3]:

```
# Checking for missing or null values
missing_values = brands_df.isnull().sum()
print(f"\nMissing values:\n{missing_values}")
```

```

Missing values:
barcode           0
category        155
categoryCode    650
name              0
topBrand        612
_id.$oid          0
cpg.$id.$oid      0
cpg.$ref          0
brandCode       234
dtype: int64
```

In [4]:

```
# Checking data types
data_types = brands_df.dtypes
print(f"\nData types:\n{data_types}")
```

```

Data types:
barcode         object
category        object
categoryCode    object
name            object
topBrand        object
_id.$oid        object
cpg.$id.$oid    object
cpg.$ref        object
brandCode       object
dtype: object
```

In [5]:

```
unique_categories = brands_df['category'].unique()

print(f"\nUnique values in 'category': {unique_categories}")
```

```

Unique values in 'category': ['Baking' 'Beverages' 'Candy & Sweets' 'Co
ndiments & Sauces'
 'Canned Goods & Soups' nan 'Magazines' 'Breakfast & Cereal'
 'Beer Wine Spirits' 'Health & Wellness' 'Beauty' 'Baby' 'Frozen'
 'Grocery' 'Snacks' 'Household' 'Personal Care' 'Dairy'
 'Cleaning & Home Improvement' 'Deli' 'Beauty & Personal Care'
 'Bread & Bakery' 'Outdoor' 'Dairy & Refrigerated']
```

In [6]:

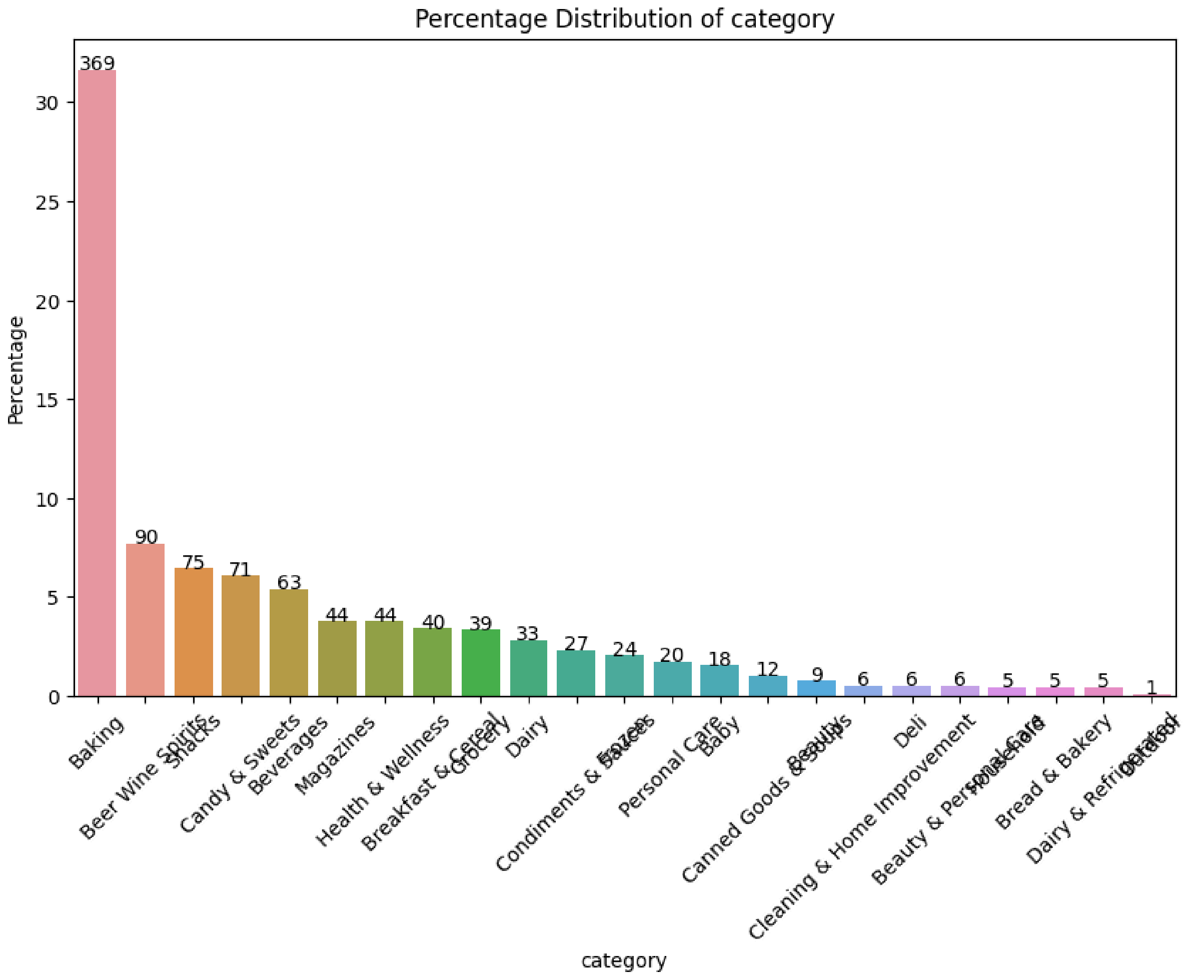
```
def plot_percentage_bar_chart(column_name, data_frame):
    total_count = len(data_frame)
    value_counts = data_frame[column_name].value_counts()
    percentages = (value_counts / total_count) * 100

    plt.figure(figsize=(10, 6))
    sns.barplot(x=percentages.index, y=percentages.values)
    plt.title(f'Percentage Distribution of {column_name}')
    plt.xlabel(column_name)
    plt.ylabel('Percentage')
    plt.xticks(rotation=45)

    for index, value in enumerate(value_counts):
        plt.text(index, percentages.values[index], f'{value}', ha='cente

    plt.show()


plot_percentage_bar_chart('category', brands_df)
```



The Data quality issues I found in the brands.json-

1. There are a lot of Missing Data in the categoryCode and topBrand.
2. There are Inconsistency in the barcode column since some barcodes do not follow the expected 12-digit numerical pattern, which suggests invalid or incorrectly formatted data

3. Baking has the higher percent of category.

In [ ]:

In [ ]:

In [ ]: